VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

DIPLOMOVÁ PRÁCE

Brno, 2019 Bc. Erika Káčerová

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV TELEKOMUNIKACÍ
DEPARTMENT OF TELECOMMUNICATIONS

ODHAD FORMANTOVÝCH KMITOČTŮ POMOCÍ STROJOVÉHO UČENÍ
ESTIMATION OF FORMANT FREQUENCIES USING MACHINE LEARNING

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. Erika Káčerová**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Jiří Mekyska, Ph.D.**
SUPERVISOR

BRNO 2019

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní obor **Audio inženýrství**
Ústav telekomunikací

Studentka: Bc. Erika Káčerová ***ID:*** 151711
Ročník: 2 ***Akademický rok:*** 2018/19

NÁZEV TÉMATU:

Odhad formantových kmitočtů pomocí strojového učení

POKYNY PRO VYPRACOVÁNÍ:

V rámci práce bude vytvořen nový algoritmus robustního odhadu formantových kmitočtů řeči, který bude využívat strojového učení v kombinaci s lineární predikcí, mapováním spektrální obálky a popisu časového průběhu. Pomocí softwaru Praat a Wavesurfer bude vytvořena referenční databáze vzorků znělé řeči (obsahující obě pohlaví a různé věkové kategorie), přičemž ke každému vzorku budou zjištěny první tři hodnoty formantových kmitočtů. Pomocí této databáze bude natrénován matematický model, který bude následně provádět odhad formantových kmitočtů automatizovaně. Model bude otestován.

DOPORUČENÁ LITERATURA:

[1] PSUTKA, Josef, et al. Mluvíme s počítačem česky. Praha: Academia, 2006. 752 s. ISBN 80-200-1309-1.

[2] KIM, CH.; KWANG-DEOK, S.; SUNG, W.: A Robust Formant Extraction Algorithm Combining Spectral Peak Picking and Root Polishing. EURASIP Journal on Advances in Signal Processing, roč. 2006, 2006: s. 1–16. ISSN 1687-6180.

Termín zadání: 1.2.2019 ***Termín odevzdání:*** 16.5.2019

Vedoucí práce: Ing. Jiří Mekyska, Ph.D.
Konzultant:

prof. Ing. Jiří Mišurec, CSc.
předseda oborové rady

ABSTRAKT

Diplomová práca sa zoberá problematikou odhadu formantových kmitočtov. V prostredí Matlab je vytvorený systém, ktorý generuje databázu referenčných hodnôt prvých troch formantových kmitočtov z nahrávok ľudskej reči. Pritom sú využité softvéry Praat a WaveSurfer(Snack). Zo zvukových súborov sú extrahované lineárne predikčné koeficienty a melovské kepstrálne koeficienty. Vytvorená databáza je použitá k trénovaniu modelu neurónovej siete. Model je v závere testovaný.

KĽÚČOVÉ SLOVÁ

formant, formantový kmitočet, LPC, Matlab, MFCC, neurónové siete, Praat, reč, spracovanie reči, strojové učenie

ABSTRACT

This Master's thesis deals with the issue of formant extraction. A system of scripts in Matlab interface is created to generate values of the first three formant frequencies from speech recordings with the use of Praat and Snack(WaveSurfer). Mel Frequency Cepstral Coefficients and Linear Predictive Coefficients are extracted from the audio files in order to be added to the database. This database is then used to train a neural network. Finally, the designed neural network is tested.

KEYWORDS

Formant, Formant Frequencies, LPC, Matlab, MFCC, Neural Networks, Praat, Speech, Speech Processing, Machine Learning

KÁČEROVÁ, Erika. *Odhad formantových kmitočtů pomocí strojového učení*. Brno, 2019, 45 s. Diplomová práca. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací. Vedúci práce: Ing. Jiří Mekyska, Ph.D.

VYHLÁSENIE

Prehlasujem, že som svoju diplomovú prácu na tému „Odhad formantových kmitočtů pomocí strojového učení“ vypracovala samostatne pod vedením vedúceho diplomovej práce, s využitím odbornej literatúry a ďalších informačných zdrojov, ktoré sú všetky citované v práci a uvedené v zozname literatúry na konci práce.

Ako autorka uvedenej diplomovej práce ďalej vyhlasujem, že v súvislosti s vytvorením tejto diplomovej práce som neporušila autorské práva tretích osôb, najmä som nezasiahla nedovoleným spôsobom do cudzích autorských práv osobnostných a/alebo majetkových a som si plne vedomá následkov porušenia ustanovenia § 11 a nasledujúcich autorského zákona Českej republiky č. 121/2000 Sb., o práve autorskom, o právach súvisiacich s právom autorským a o zmene niektorých zákonov (autorský zákon), v znení neskorších predpisov, vrátane možných trestnoprávnych dôsledkov vyplývajúcich z ustanovenia časti druhej, hlavy VI. diel 4 Trestného zákoníka Českej republiky č. 40/2009 Sb.

Brno ....

podpis autorky

POĎAKOVANIE

V prvom rade ďakujem pánovi Ing. Jiřímu Mekyskovi, Ph.D. za jeho ochotu pri odbornom vedení tejto práce, za prínosné rady a cenné pripomienky, bez ktorých by táto práca nevznikla. Ďakujem aj svojej rodine, za mnohostrannú podporu a trpezlivosť počas všetkých mojich študijných rokov. Najväčšie ďakujem však patrí môjmu priateľovi, ktorý mi počas celého štúdia nedovolil to vzdať.

Brno . . . .

podpis autorky

Obsah

Zoznam obrázkov

Zoznam tabuliek

Úvod

Táto diplomová práca sa zaoberá problematikou odhadu formantových kmitočtov. Formanty vznikajú dôsledkom prechodu prúdu vzduchu rezonančnými dutinami v ľudskom tele pri tvorbe reči. Tieto rezonancie zvýrazňujú časti spektra v okolí určitých kmitočtov.

Znalosť polohy formantových kmitočtov sa zaraďuje medzi dôležité parametre rečových signálov. Oblasť využitia tejto znalosti je široká. Z formantových charakteristík je možné čerpať pri modelovaní filtru hlasového traktu. Na princípe formantovej syntézy boli založené prvé elektrické syntezátory už na začiatku 20. storočia. Dnes sa uplatnenie ponúka v rámci identifikácie pohlavia a veku hovoriaceho alebo pri pozorovaní zmeny zdravotného stavu pacienta.

V súčastnosti je známych viacero postupov odhadu formantových kmitočtov. Žiadna z nich však nie je presná a bezproblémová. Zaužívané metódy často narážajú na komplikácie, ktoré spôsobujú chyby vo výsledkoch. Momentálne neexistuje nástroj, ktorý by bol schopný automaticky určiť hodnoty formantových kmitočtov bez potreby zadania akéhokoľvek vstupného parametru. Ponúka sa myšlienka naučiť rozpoznať formantové kmitočty model strojového učenia.

Strojové učenie dnes ponúka rôznorodé možnosti vyhodnocovania a predikcie znalosti s veľkým potenciálom. Jeho metódy poskytujú širokú škálu algoritmov, ktoré dokážu nájsť uplatnenie takmer pri každom, aj netriviálnom probléme. Naskytuje sa teda možnosť navrhnúť a pomocou referenčnej databáze natrénovať určitý model strojového učenia, ktorý bude mať príležitosť vykonávať odhad formantových kmitočtov úplne automatizovane.

Jedným z dvoch hlavných cieľov tejto diplomovej práce je vytvorenie skriptu v prostredí Matlab, ktorý umožní jednoduché vygenerovanie referenčnej databáze hodnôt formantových kmitočtov. Vytvorenie tejto referenčnej databáze je nevyhnuté pre následné trénovanie a testovanie matematického modelu strojového učenia.

Druhým hlavným cieľom je navrhnutie, praktické vytvorenie a natrénovanie vybraného modelu strojového učenia. V závere práce prebehne testovanie vytvoreného modelu a vyhodnotenie jeho presnosti určovania formantových kmitočtov.

Kapitola 1 približuje princípy tvorenia reči, vysvetľuje vznik javu formantov a pojednáva o problematike odhadu formantových kmitočtov. Kapitola 2 načrtáva príležitosti strojového učenia a predkladá možnosti použitia jeho algoritmov. Nasledujúca kapitola 3 sa venuje praktickej časti prípravy trénovacej databáze. Najskôr je predložené zhrnutie možností dostupných korpusov hovorenej českej reči, z ktorých je jeden vybraný k ďalšiemu spracovaniu. Následne je popísaný postup naprogramovania skriptu pre vytvorenie databáze, spojený s vyhotovením referenčnej databáze

hodnôt formantových kmitočtov. Pomocou vytvorenej databáze je natrénovaný model neurónovej siete.

1 Formantové kmitočty

Formantové kmitočty patria k jedným z najdôležitejších parametrov rečových signálov. Formantová analýza nachádza uplatnenie pri rôznych aplikáciach ako rozpoznanie reči, charakterizácia rečníka [1], ale využitie sa ponúka aj pre účely identifikácie rečníkov, ich pohlavia alebo veku. Táto kapitola pojednáva práve o problematike javu vzniku formantových kmitočtov a odhadu ich hodnôt z akustického signálu.

1.1 Tvorba reči

Tvorbou reči sa z fyziologického hľadiska zaoberá jazykovedná disciplína nazývaná *fonetika*. Fonetika skúma a popisuje mechanizmus vzniku reči, jej akustickú stavbu ale aj vnímanie reči sluchom. Jej základnou jednotkou slúžiacou k popisu jazyka je hláska, ktorá odpovedá minimálnej zvukovej jednotke reči, patriacej určitému jazyku.

Ďalšou dôležitou disciplínou pri štúdiu reči je *fonológia*. Zaoberá sa funkciou hlások a skúma významotvorné rozdiely medzi nimi. [2]

Za samotnú tvorbu reči u človeka sú zodpovedné rečové orgány hlasového traktu, ktoré tvoria *rečové ústrojenstvo*, viď obr. 1.1. Ten môžeme rozdeliť na 3 časti:

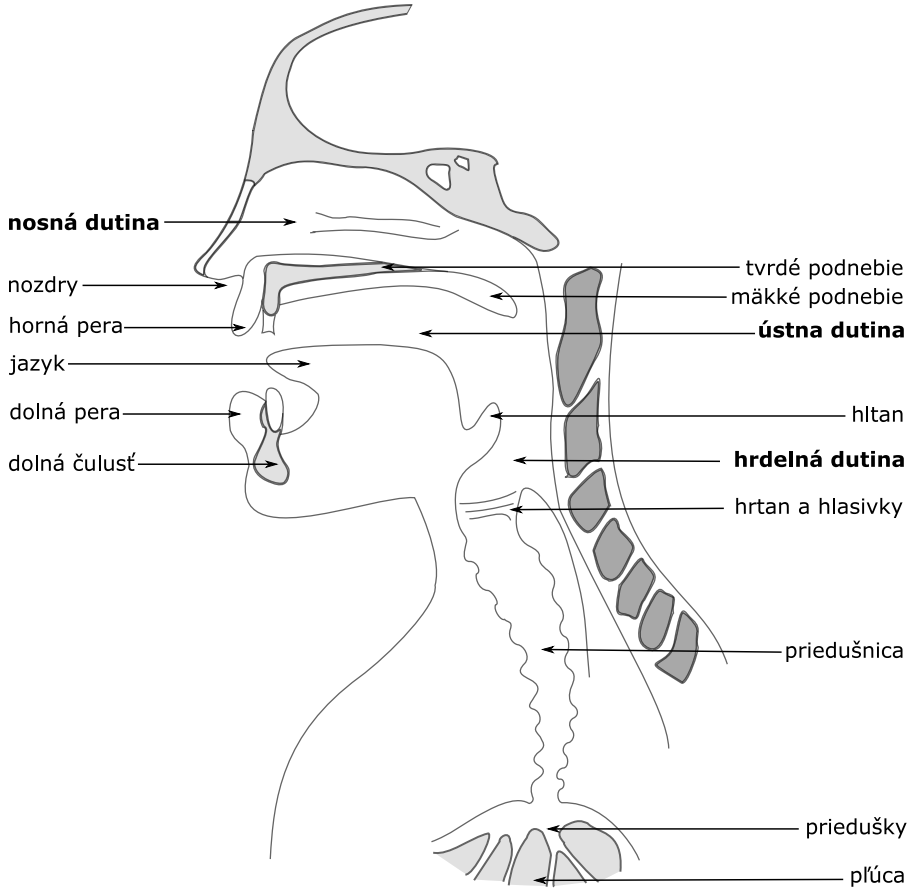
- dýchacie (respiračné) ústrojenstvo,
- hlasové (fonačné) ústrojenstvo,
- modifikujúce (artikulačné) ústrojenstvo.

1.1.1 Dýchacie ústrojenstvo

Už podľa názvu primárne slúži k jednej zo základných životných funkcii – dýchaniu. Pri tvorbe reči však výdych zabezpečuje pohyb základného materiálu, potrebného pre vznik akéhokoľvek zvuku – vzduchu.

Práca dýchacích orgánov sa odlišuje pri fyziologickom dýchaní a pri tvorbe reči. Rozdiel môžeme pozorovať v rytme dýchania, ale aj množstve vzduchu s ktorým orgány pracujú. Pri klasickom dýchaní je nádych a výdych v pomere 2:3, pričom pri tvorbe reči sa čas nádychu výrazne skracuje. Naopak sa predlžuje čas pre výdych, pri ktorom zvyčajne dochádza k tvorbe reči[1]. Objem vzduchu, ktorý vdychujeme prirodzene pri dýchaní odpovedá približne 0,5 l, pri rozprávaní sa zvyšuje až na

[1] Tvorba hlások pri nádychu nie je síce obvyklá, ale rozhodne nie je nemožná. V cudzích jazykoch sa stretneme s niektorými hláskami, ktoré vznikajú zásadne vdychom vzduchu. Rovnako v češtine môžeme pozorovať tvorbu zvukov pri nádychu, napríklad pri niektorých citoslovciach ako *hí* - zvuk prekvapivej reakcie



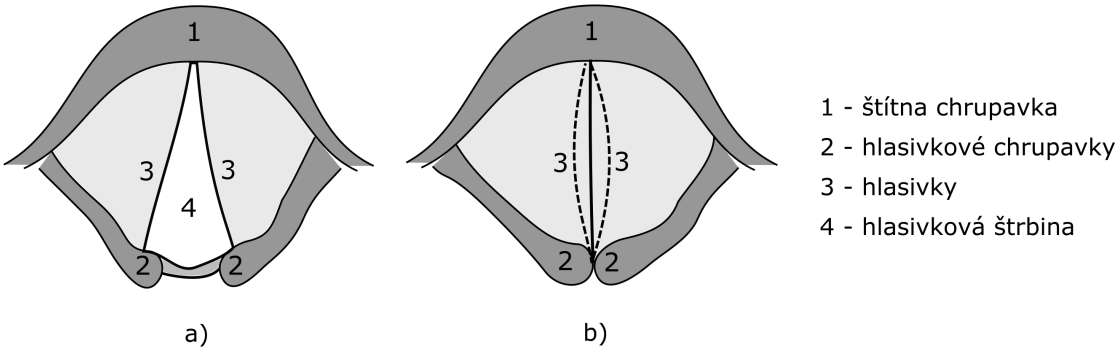
Obr. 1.1: Hlasový trakt človeka.

trojnásobok. Vzhľadom na to, že pri tvorbe reči je potrebné inhalovať za kratší čas väčší objem vzduchu, sú do nádychu prirodzene zapojované aj ústa. [2]

Práca pľúc, dýchacích ciest a bránice teda zabezpečuje prúdenie základného zdroja energie pre reč hlasovým traktom.

1.1.2 Hlasové ústrojenstvo

Hlasové ústrojenstvo sa nachádza v hrtane. Práve tu vzniká základný ľudský hlas, ktorého tvorbu majú na starosti *hlasivky*. Hlasivky sú dve ostré sliznicové riasy, napnuté medzi štítnou chrupavkou a párom hlasivkových chrupaviek (obr. 1.2). Štítna chrupavka je pevná a konce hlasiviek, ktoré sú s ňou spojené sa preto vždy dotýkajú. Naopak hlasivkové chrupavky sú čiastočne pohyblivé. Dokážu sa otáčať, odďalovať alebo meniť sklon [2]. Týmto pohybom sa druhý koniec hlasiviek, ktorý je pripojený na tieto chrupavky dokáže oddialiť a vytvoriť tak hlasivkovú štrbinu (glottis) o veľkosti až 8 mm. Táto štrbina je využívaná keď sa netvorí reč. Hlasivky sú v kľudnom stave, vzduch cez ne môže prúdiť a človek voľne dýcha [3].



Obr. 1.2: Schéma hlasiviek: a) pri kľudovom postavení (dýchanie), b) pri fonačnom postavení (kmitanie).

Naopak pri zámere tvorby zvuku sa hlasivkové chrupavky priblížia, čo uzavrie štrbinu medzi hlasivkami. Hovoríme o *fonačnom postavení* hlasiviek. Vzduch prúdiaci hrtanom pri jeho prechode v tomto prípade hlasivky rozkmitáva, čím dochádza k tvorbe základného tónu hlasu. Frekvencia kmitania tohoto tónu býva označovaná ako F_0 a nazýva sa *fundamentálny kmitočet* alebo *kmitočet základného (hlasivkového) tónu*. Jeho obrátenú hodnotu $T_0 = 1/F_0$ nazývame *perióda základného (hlasivkového) tónu* (pitch). Udáva sa, že pri reči sa hodnota F_0 pohybuje v rozmedzí asi $60 - 400\,\mathrm{Hz}$. Konkrétny rozsah sa líši u žien, mužov a detí, pri speve však F_0 môže až prevýšiť hodnotu $1000\,\mathrm{Hz}$. [3]

Výška základného tónu hlasu vychádza z anatómie hlasiviek, význam má však predovšetkým ich dĺžka. U mužov pozorujeme väčšiu dĺžku hlasiviek (typicky cca $15\,\mathrm{mm}$ [3], bas až 24–$25\,\mathrm{mm}$ [2]), čo odpovedá najnižšiemu základnému tónu hlasu. Ženy naopak majú hlasivky kratšie (typicky cca $13\,\mathrm{mm}$ [3], soprán 14–$19\,\mathrm{mm}$ [2]). Kratšie hlasivky kmitajú rýchlejšie, čo spôsobuje vyšší základný tón hlasu.

Postavenie hlasiviek a typ kmitania sa líši pri tvorbe rozličných skupín hlások. Pri znelých zvukoch (samohlásky a znelé spoluhlásky) je charakteristické fonačné postavenie, ktoré má za následok vznik hlasivkového tónu. U samohlások je štrbina úplne uzavrená, viď obr. 1.2. Hlasivky sú tesne zblížené po celej dĺžke. Kmitanie je tak pravidelné a má za následok typické kvázi-periodický charakter, pri ktorom je možné pozorovať tónovú štruktúru. Kmitanie pozorujeme aj pri tvorbe znelých spoluhlások, napätie hlasiviek je ale menšie. Charakteristika kmitania je menej pravidelná a už nie je čisto tónová.

Neznelé spoluhlásky sú tvorené pri kľudovom postavení hlasiviek. Preto neobsahujú F_0 a typický zvuk vzniká až v nad-hrtanových dutinách. Pri tvorbe explo-

zívnych spoluhlások (napr. [p], [t], [ť]) sú tesne priblížené. Poloha takmer zhodná s dýchaním, teda úplne rozovretá, vytvára spoluhlásky so šumovou charakteristikou (napr. [f], [s], [š], [ch]). V jedinečnom postavení sa hlasivky nachádzajú pri tvorbe českého [h] – sú priblížené v blanitej časti a kmitajú pomalšie.

Hlas ktorý produkujú hlasivky ešte nemá znenie individuálneho rečníka. Táto charakteristika vzniká až pri prechode nad-hrtanovými priestormi a je tvorená rezonanciami v dutinách týchto priestorov. Až tu vzniká jedinečná farba každého ľudského hlasu. [2, 3]

1.1.3 Artikulačné ústrojenstvo

Artikulačné ústrojenstvo umožňuje vytvárať veľké množstvo rôznych zvukov, preto má veľký podiel na tvorbe reči. Nachádza sa nad hrtanom a je tvorené tromi rezonančnými dutinami (obr. 1.1):

- hrdelná dutina,
- nosná dutina,
- ústna dutina.

Hrdelná dutina sa rozprestiera priamo nad hlasivkami. Na tvorbe charakteristiky hlasu jednotlivca sa účastní ako rezonančný priestor. Jeho tvar a objem je menený iba pasívne pohybom koreňu jazyka a hrtanu, ktorý môže meniť celkovú dĺžku hlasového traktu. Zároveň však aktuálne rozmery tohoto priestoru ovplyvňuje aj činnosť svalov hrdla a krku. Psychický stav hovoriaceho má veľký vplyv na uvoľnenie, alebo prípadné napätie týchto svalov. Napríklad nervozita hovoriaceho spôsobuje stiahnutie krčných svalov, tým zmení rezonanciu hrdelnej dutiny, čo má za následok zmenu farby hlasu, prípadne rozochvenie, ktoré je poslucháč schopný vnímať a to aj v jazyku, ktorému nerozumie.

Nosná dutina je rovnako rezonančným priestorom. Uplatňuje sa však pri tvorbe iba niektorých hlások, najviac pri tzv. *nazálach* – nosných hláskach – v češtine [m], [n], [ň]. Pohyb mäkkého podnebia pri vyslovovaní ostatných hlások spôsobuje čiastočné alebo úplne uzavretie priechodu medzi nosnou dutinou. V tomto prípade je táto dutina od vzduchu prúdiaceho hlasovým traktom oddelená a nepodieľa sa na tvorbe charakteristiky hlasu. Toto oddelenie spôsobuje mäkké podnebie. Je to sval, v ktorom sa pri artikulácií nachádza určité napätie. Toto napätie nie je pri vyslovovaní všetkých hlások rovnaké, preto sa aj pri niektorých nenazálnych hláskach môže nosná dutina prejaviť čiastočne (najviac pri hláske [a]).

Ústna dutina Procesy prebiehajúce v ústnej dutine sú pri tvorbe hlások zaujímavejšie ako v predchádzajúcich priestoroch. Má najväčší význam pre výslovnosť hlások, pretože v týchto priestoroch dochádza k diferenciácii takmer všetkých zvukov. Tu sa nachádzajú najvýznamnejšie artikulátory, ktorými hovoriaci pri rozprávaní pohybuje. Vedomý pohyb čeľustí, jazyka a pier spôsobuje zmenu tvaru a objemu tejto dutiny. [2]

1.1.4 Tónová štruktúra reči

Konkrétna poloha artikulačných orgánov teda vytvára jedinečnú kombináciu objemov rezonančných dutín. Rezonancia prúdu vzduchu týmito dutinami spoločne s určitým typom chvenia hlasiviek vytvára jedinečný zvuk, odpovedajúci zvukovému zastúpeniu jednotlivej hlásky.

Rezonancie v hrdelnej, ústnej a prípadne nosnej dutine vytvárajú zmenu rozloženia akustickej energie vo vznikajúcom signále. Tvar a objem týchto dutín v konkrétnom postavení orgánov hlasového traktu vytvára sústredenie akustickej energie okolo určitých kmitočtov. To má za následok sústredenie akustickej energie v okolí určitých kmitočtov v spektre rečového signálu. Tieto oblasti zváraznenia sú označované ako *formanty* a kmitočty, v ktorých sa vyskytujú ako *formantové kmitočty*. [3]

1.2 Vlastnosti hlások

České hlásky sú na základe ich akustickej charakteristiky jednoducho rozdeliteľné do dvoch skupín[2]:

- **Vokály** (*samohlásky*) – charakteristické budením kvaziperiodickým signálom so základným tónom, ktorého spektrum je ďalej upravované kmitočtovou charakteristikou rezonančných dutín hlasového traktu.
- **Konsonanty** (*spoluhlásky*) – typickou základnou zložkou je šum, či už samostatný u tzv. skutočných *neznelých spoluhlások*, alebo doplnený o tónovú zložku pri *znelých spoluhláskach* (mierne odlišnú od čistej samohláskovej tónovej štruktúry).

Samohlásky teda vokály, sú tvorené jasnou tónovou charakteristikou. Hlasivky sú vo fonačnom postavení, blízko u seba a napnuté, takže pri prechode vzduchu kmitajú kvázi-periodickými pulzmi. Tento signál je ďalej modifikovaný rezonanciami

[2]Toto delenie je síce teoreticky zreteľne jasné, prakticky však nie je možné každú spoluhlásku zaradiť do jednej z týchto kategórií. Existujú výnimky, ako napríklad hláska j, ktorého budiaci signál je tvorený zmesou základného tónu a šumu, ale jeho veľkosť je malá.

nad-hrtanových dutín. Mäkké podnebie je pri tvorbe všetkých samohlások zdvihnuté. Priechod do nosnej dutiny je tým uzavretý a k rezonancii v tejto dutine nedochádza. Výrazne sa uplatňuje činnosť jazyka, ktorého poloha mení tvar a objem dutiny ústnej. Túto polohu pri vyslovovaní jednotlivých vokálov je možné vyčítať z tabuľky 1.1. Pohyb pier je znateľný najmä pri samohláskach s ich extrémnou pozíciou – maximálne napätie pri hláske [i], naopak maximálne zaokrúhlenie pri hláske [u]. [4]

Tab. 1.1: Hellwagov vokalický trojuholník českých hlások [2]

	Rozdelenie podľa pohybu v horizontálnom smere		
Rozdelenie podľa pohybu vo vertikálnom smere	predné (anteriorne)	stredné (centrálne)	zadné (posteriorne)
zavreté (vysoké)	i		u
stredové	e		o
otvorené (nízke)		a	

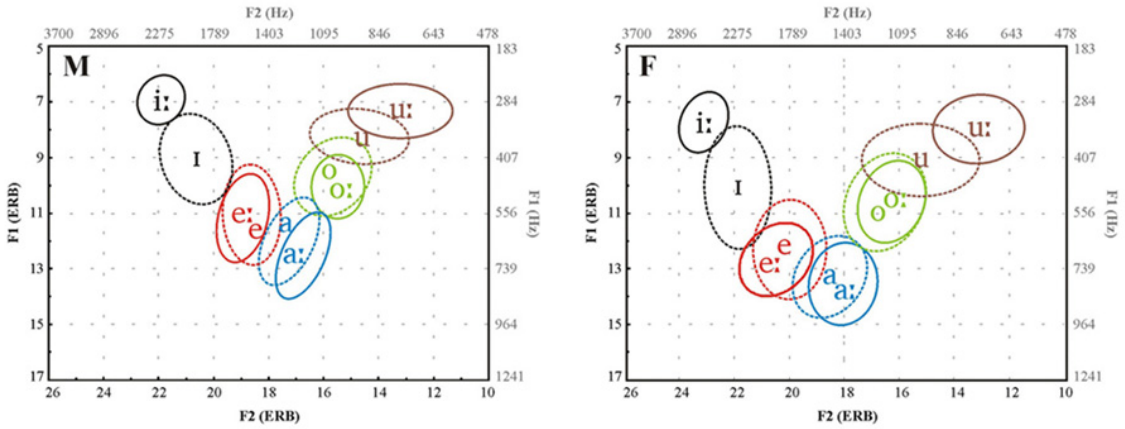
Fundamentálny hlas je pre všetky vokály zhodný. K diferenciácií jednotlivých samohlások dochádza až pri priechode rezonančným priestorom. Rezonátory sa teda podieľajú na výslovnosti každého vokálu. Dochádza k vzniku *formantov* – typických oblastí zvýraznenia spektra rečového signálu. Formanty sú označované podľa ich poradia v spektre od prvého ako F_1, cez druhý F_2 až po N-ty F_N, pričom označenie F_0 je rezervované pre základný tón hlasu. K vytvoreniu a následnej identifikácii jednotlivých hlások sú dôležité najmä prvé dva formanty. Ľudská reč ich síce obsahuje viac (uvádza sa až 6), ale ako je možné vidieť v tabuľke č. 1.2, rozsah hodnôt sa už u tretieho formantu pri rôznych vokáloch výrazne nelíši, pre niektoré je dokonca zhodný. Poloha vyšších formantov síce nemá na určenie samohlások hodnotný význam, vytvára však jedinečnú farbu hlasu jedinca. [2]

Z tabuľky 1.2 je teda zreteľné, že pre rozpoznanie jednotlivých vokálov sú dôležité formanty F_1 a F_2. Rozdiel medzi samohláskami tvorí jedinečná kombinácia rezonancií na formantových kmitočtoch. K identifikácií konkrétnej hlásky preto nie sú dôležité absolútne hodnoty týchto formantov, ale ich vzájomný pomer, teda pomer medzi kmitočtom prvého formantu F_1 a druhého formantu F_2.

V roku 2017 vznikol na Fonetickom ústave Filozofickej fakulty Univerzity Karlovej v Prahe výskum, skúmajúci polohu formantov mladých dospelých hovoriacich. Výskum analyzoval hodnoty samohláskových formantov od 75 vysokoškolských študentov, zahŕňajúci obe pohlavia. Na obrázku 1.3 sú viditeľné výsledky tohoto výskumu zvlášť pre ženy a mužov. [6]

Tab. 1.2: Rozsah hodnôt prvých troch formantových kmitočtov českých vokálov v Hz. [5, 3]

Samohlásky	F_1 pásmo	F_2 pásmo	F_3 pásmo
i, í	300 – 500	2000 – 2800	2600 – 3500
e, é	480 – 700	1560 – 2100	2500 – 3000
a, á	700 – 1100	1100 – 1500	2500 – 3000
o, ó	500 – 700	850 – 1200	2500 – 3000
u, ú	300 – 500	600 – 1000	2400 – 2900



Obr. 1.3: Referenčné hodnoty formantových kmitočtov u mužov (M) a žien (F) z [6]

1.3 Odhad formantových kmitočtov

Väčšina zaužívaných postupov identifikácie formantových kmitočtov využíva buď implicitne, alebo explicitne spektrálnej obálky, pretože informácie o formantoch sú dokázateľne obsiahnuté v jej tvare. Tieto postupy teda pracujú v kmitočtovej oblasti. Základom je analýza spektrálnej obálky, stanovenej metódou lineárne predikčnej analýzy (LPC – Linear Predictive Coding). Produktom tejto analýzy sú tzv. *lineárne predikčné koeficienty* – LPC koeficienty. Prenosová funkcia modelu hlasového traktu má tvar

$$H_v(z) = \frac{G_v}{A(z)} = \frac{G_v}{\sum_{k=0}^{p} a_k z^{-k}} \ , \tag{1.1}$$

kde G_v je zosilnenie vokálneho traktu a p označuje rád lineárnej predikcie. Koeficienty a_k označujú práve LPC koeficienty, modelované lineárnou predikčnou analýzou, pričom vždy platí $a_0 = 1$ [4].

Zdroj [3] uvádza dve základné možnosti postupu tejto varianty:

1. výpočet pólov prenosovej funkcie a
2. vyhľadanie vrcholov spektrálnej obálky.

Metóda výpočtu pólov prenosovej funkcie zisťuje póly funkcie $H_v(z)$, teda hľadá korene polynómu $A(z)$. Tie sú získané riešením rovnice

$$A(z) = \sum_{k=0}^{p} a_k z^{-k} = 0. \tag{1.2}$$

Jedná sa o rovnicu rádu p s reálnymi koeficientmi. Riešením sú prevažne páry komplexne združených koreňov [3].

Ak uvažujeme jednu dvojicu komplexne združených koreňov, odpovedajúci formantový kmitočet F_i je možné vyjadriť pomocou vzťahu

$$F_i = \frac{\omega_i}{2\pi} = \frac{\arg(z_i)}{2\pi T} \quad [\mathrm{Hz}], \tag{1.3}$$

následne šírku pásma B_i pre pokles modelu spektrálnej obálky o 3 dB vzťahom

$$B_i = -\frac{\ln|z_i|}{\pi T} \quad [\mathrm{Hz}], \tag{1.4}$$

kde v obidvoch rovniciach T zastupuje vzorkovaciu periódu akustického signálu [1].

Metóda vyhľadávania vrcholov spektrálnej obálky rovnako využíva tvaru spektrálnej obálky. Formanty nájdené prehľadaným tejto obálky, zistením bodov lokálnych maxím a následnou identifikáciou hodnôt kmitočtov pre tieto maximá. K zisteným formantovým kmitočtom F_i je možné rovnako doplniť šírku pásma B_i poklesu o 3 dB. Pri použití tejto metódy je vhodné z kandidátov na formanty vylúčiť ktorých šírka pásma prevyšuje 500 Hz, pretože šírka pásma reálnych formantov nedosahuje tieto hodnoty. [3]

2 Strojové učenie

Strojové učenie (anglicky Machine Learning) patrí pod odbor *umelej inteligencie*. Cieľom umelej inteligencie je napodobniť myslenie ľudského mozgu. Snaží sa teda prinútiť stroje k takému správaniu, ktoré by u človeka vyžadovalo potrebu inteligencie. [7]

Strojové učenie je súbor algoritmov, ktoré umožňujú umelým objektom proces *učenia*. Produktom týchto algoritmov je *model*, naučený určitému správaniu. Učenie v tomto prípade predstavuje postup automatického zlepšovania na základe skúseností. Stroje pri tom majú byť schopné sformulovať popis žiadaného pojmu na základe určitých charakteristických vlastností [8].

Neoddeliteľnou súčasťou vytvorenia a vylepšovania modelu sú metódy z ďalších vedných odborov, *pravdepodobnosti a štatistiky* a *optimalizácie*.

Štatistické postupy sú v rámci strojového učenia využívané vo fáze pred-spracovania dát, ale aj vyhodnocovania modelu. Algoritmy a modely strojového učenia sa dajú prirovnať k termínu *hypotéza*, pretože model je vo svojej podstate hypotézou vzťahov a väzieb medzi vstupnými a výstupnými dátami modelovaného procesu. Niektoré algoritmy dokonca priamo vychádzajú z postupov štatistických metód.

Pri trénovaní modelu zohráva dôležitú úlohu aj optimalizácia. Využívané sú pritom algoritmy slúžiace k identifikácii extrému funkcie. Optimalizácia pracuje s modelom ako s funkciou, na ktorej hľadá minimum (napr. metóda hladania čo najnižšej *Mean Square Error* (MSE)), prípadne maximum (Maximum Likelihood – najväčšiu vieruhodnosti). Cieľom je nadstaviť model tak, aby sa minimalizovala chyba predikcie, teda odchýlka medzi rozdielom výstupných hodnôt modelu a požadovanými výstupnými hodnotami. [7]

Ná základe prítomnosti informácie o výstupe existujú dva typy algoritmov:

- *Supervised learning* – učenie s učiteľom, kde tréningové dáta obsahujú hodnoty požadovaných výstupov,
- *Unsupervised learning* – učenie bez učiteľa, tréningový dataset neobsahuje očakávané výstupy.

Z pohľadu druhu výstupu sú rozlišované tieto typy predikcie:

1. *Klasifikácia* rieši priradenie daného prvku k jednej z klasifikačných skupín. Počet klasifikovaných skupín odpovedá počtu výstupov. Ideálne by mal každý z prvkov trénovacej skupiny patrí práve do jednej z klasifikačných skupín, takže má na výstupoch nuly a jednu jednotku. Predikované hodnoty neznámych prvkov sa pohybujú v rozsahu 0 až 1 a prvok je zaradený do skupiny s najvyššou hodnotou.
2. *Regresia* naopak na každom z výstupov predikuje hodnotu v ľubovolnom rozsahu. Trénovacie dáta nemajú na výstupoch iba jednotky a nule ale čísla, ktoré

reálne niečo predstavujú.

V následujúcich sekciách budú stručne popísané niektoré konkrétne metódy strojového učenia, ktorých modely bude možné natrénovať a testovať v rámci nadväzujúcej diplomovej práce.

2.1 Metóda *k*-najbližších susedov

Základom tejto metódy je analógia modelu n-dimenzionálneho priestoru. Osi priestoru predstavujú vstupné príznaky, n sa teda rovná počtu príznakov. Jednotlivé prvky trénovacej množiny tvoria body v priestore. Trénovanie tejto metódy je teda jednoduché, spočíva iba v uložení trénovacích prvkov do priestoru. Každý bod, reprezentujúci trénovací prvok má presne určenú svoju polohu hodnotami svojich príznakov.

V prípade klasifikácie sa o novom prvku rozhoduje na základe vzdialeností od už uložených bodov. Existujú viaceré možnosti výberu distančnej funkcie, najčastejší je však Euklidovský vzorec:

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \tag{2.1}$$

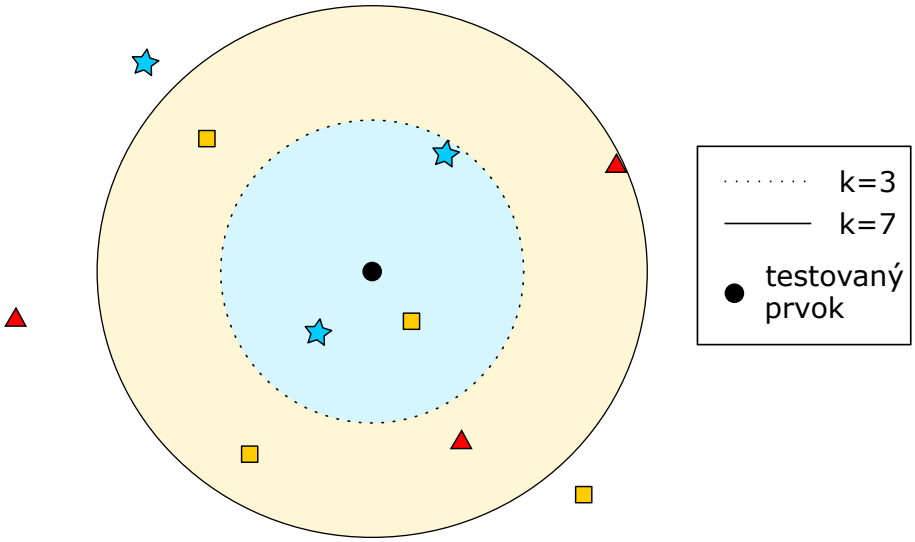
kde d je vzdialenosť bodov X a Y, x_i je hodnota i-teho príznaku prvku X, obdobne y_i.

Na obrázku 2.1 možno pochopiť rozdiel výsledku klasifikácie pri rôznych hodnotách k. Testovaný prvok je vždy zaradený do triedy, ktorá prevláda medzi jeho k-najbližšími susedmi. Pri voľbe k=3 bude teda zaradený do triedy hviezdička, no pri voľbe k=7 to už bude trieda štvorec. Táto metóda je citlivá na vhodnú voľbu hodnoty k.

2.2 Neuronové siete

Neurónové siete sú jednou z tradičných metód strojového učenia. Počiatky realizácie tejto metódy vznikali už v 40. rokoch 20 storočia. Umelé neuronónové siete fungujú podľa vzoru biologických nervových systémov. Základnou jednotkou nervového systému je neurón, teda bunka, ktorá sa špecializuje na získavanie, spracovanie a ukladanie informácií. Vďaka vzájomnému prepojeniu jednotlivých neurónov, dochádza medzi nimi k prenosu informácie.

Obdobne je umelá neurónová sieť zložená z jednotlivých neurónov, usporiadaných do vrstiev. V každom modele sa nachádza jedna vrstva vstupných neurónov,



Obr. 2.1: Princíp klasifikácie algoritmu k-NN.

ktoré očakávajú na vstupe prvky trénovacích dát. Posledná vrstva je vrstvou výstupných neurónov, kde dochádza k vyvodeniu výsledku. Ďalšie neuróny sú usporiadané do jednej, alebo viacerých vrstiev medzi týmito dvomi. Neuróny sú medzi jednotlivými vrstvami prepojené a k učeniu modelu dochádza nadstavovaním váh týchto prepojení.

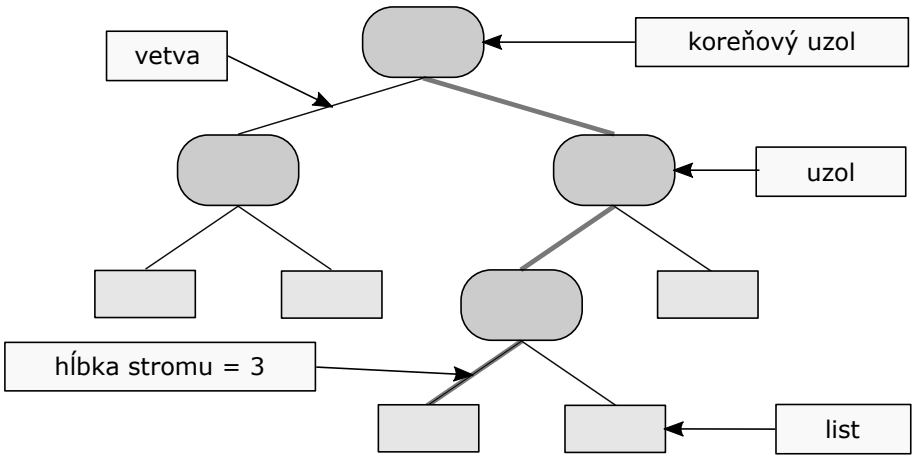
Výhodou neurónových sietí je možnosť paralelného spracovávania mnohými prvkami. [9]

2.3 Rozhodovacie stromy

Metóda rozhodovacích stromov (ang. Decision Trees) je jednoduchšie pochopiteľná ako predošlá metóda. Význačná je hierarchická štruktúra, ktorá umožňuje nájdenie a uloženie znalosti. Tento systém je možné znázorniť pochopiteľným grafickým výstupom (viď obr. 2.2), z ktorého je možné postup k výsledku jednoducho interpretovať. Základom každého rozhodovacieho stromu sú uzly, vetvy a listy, rovnako ako u stromov v prírode.

Trénovacie dáta vstúpia do uzlu, kde sa na základe podmienky rozhodne o ich ďalšom postupe. Ak sú splnené ukončovacie podmienky, na ich základe je ďalšieho uzlu vytvorený list a daný prvok trénovacej množiny je klasifikovaný. V prípade, že nedôjde k splneniu ukončovacej podmienky, dochádza k ďalšiemu vetveniu, kde sa hľadá nová, čo najvhodnejšia podmienka, aby rozvetvenie spôsobilo čo najväčší informačný zisk. Tento proces sa v každom kroku opakuje s cieľom nájdenia najefektívnejšieho postupu vetvenia. Hĺbka stromu označuje počet vetiev od koreňového uzla k najvzdialenejšiemu listu. [7]

Jednoduchosť rozhodovacieho stromu umožňuje použitie viacerých jednotlivých stromov zároveň. Táto technika býva v anglických zdrojoch označovaná ako Decision Tree Ensemble – súbor rozhodovacích stromov. Využívajú ju prepracovanejšie algoritmy ako napríklad metóda náhodných lesov (Random Forest) alebo algoritmus XGBoost.



Obr. 2.2: Príklad rozhodovacieho stromu.

XGBoost je open-source balíček implementujúci systém strojového učenia. Za pôvodného tvorcu myšlienky býva označovaný Tianqi Chen, na celkovom projekte sa podieľalo veľa ďalších vývojárov. V súčasnej dobe je veľmi populárny a projekty používajúce tento algoritmu získavajú umiestnenia na popredných rebríčkoch v medzinárodných súťažiac strojového učenia[1].

Skratka XGBoost stojí pre Extreme Gradient Boosting. Základ algoritmu je teda postavený na metóde boostovania gradientu, ktorý sa typicky využíva pri súboroch rozhodovacích stromov. XGBoost sľubuje efektívnosť, flexibilnosť a jednoduchú prenosnosť. Funguje na širokej škále rozhraní ako C++, Python, R, Java ale aj z príkazového riadku. [10]

[1]balíček spolu so zoznamom aktuálnych úspechov dostupný na <`https://github.com/dmlc/xgboost`>

3 Príprava referenčnej databáze

Ako bolo vysvetlené v kapitole 2, k vytvoreniu matematického modelu strojového učenia je potrebná databáza trénovacích prvkov. V tejto kapitole bude rozpracovaný postup prípravy tejto databáze od výberu vhodného rečového korpusu až po zápis získaných dát do súboru.

3.1 Výber rečového korpusu

Jazykový korpus je rozsiahly súbor textov, ktoré môžu mať písanú alebo hovorenú podobu, prevedený do elektronickej podoby. Cieľom budovania korpusov je účel záznamu a modelu pre pozorovanie jazyka. Korpusy tradične vznikajú so zameraním využitia v lingvistike a tomuto účelu je aj podriadená ich výstavba [11]. Keďže k vytvoreniu trénovacej databáze pre túto prácu je potrebný akustický materiál, bude táto časť skúmať hovorené korpusy v českom jazyku.

3.1.1 Český národný korpus

Projekt Český národný korpus na svojich webových stránkach sprístupňuje radu rôznych korpusov v elektronickej forme pre účely výuky a výskumu. Okrem korpusov v písanej forme obsahuje nasledujúce hovorené korpusy.

ORAL je tvorený tromi korpusmi: ORAL2006, ORAL2008 a ORAL2013, pričom prvé dva obsahujú nahrávky iba z oblasti Čiech, posledný aj z Moravy a Slezka. Korpus ORAL2013 ako jediný z nich ponúka prístup aj k zvukovým stopám nahrávok. Tie sú voľne prístupné k akademickému použitiu po prihlásení z databáze webovej stránky projektu LINDAT Univerzity Karlovy [12].

ORTOFON je nasledovníkom korpusu ORAL. Je vyvážený k pohlaviu, veku, vzdelaniu a oblasti pôvodu rečníka. Nahrávky sú rovnako dostupné pre akademické účely a po prihlásení zo zdroja [12]. ORAL však obsahuje nahrávky formou dialógu s pomerne veľkým šumom v pozadí.

DIALEKT je korpus sústreďujúci sa na zachytenie nárečového materiálu. Obsahuje tradičné teritoriálne dialekty v rámci územia celej Českej republiky.

Pražský mluvený korpus – PMK je historicky prvým hovoreným korpusom českého jazyka. Pôvodné magnetofónne nahrávky pochádzajú z rokov 1988 – 1996 z Prahy a okolia boli postupne prepisované do počítača.

Brněnský mluvený korpus – BMK obsahuje nahrávky hovorenej češtiny z oblasti Moravy z rokov 1994 – 1999.

Samostatné zvukové súbory posledných troch korpusov však nie sú dostupné k stiahnutiu. [11]

3.1.2 BUT – CZAS

Korpus vytvorený v roku 2018 na Fakulte elektrotechniky a komunikačných technológií Vysokého učení technického v Brně. Obsahuje nahrávky čítaného textu 18 rečníkov (9 mužov a 9 žien) rôznych vekových kategórií. Náhravky boli vytvorené v bezodrazovej komore vo vysokej kvalite (48 kHz, 24 bitov). Všetky dáta sú voľne dostupné k stiahnutiu na webovej stránke projektu. [13]

3.1.3 UWB

Skratkou UWB označuje svoje korpusy Katedra Kybernetiky Západočeskej univerzity v Plzni.

UWB_S01 je korpus čítaných textov z roku 2004 od 100 rečníkov (64 mužov a 36 žien). Texty k nahrávkam pochádzali z troch českých denníkov. Nahrávané bolo na 2 mikrofóny, jeden zachytával kvalitne hlas rečníka, druhý spoločne aj okolitý šum. [14]

UWB-05-HSCAVC je jedinou českou multimediálnou databázou spojitej reči, vytvorenou pre účely audiovizuálneho rozpoznávania reči. Licencia tohoto korpusu nie je voľná a jej akademická nekomerčná verzia stojí 550 €. [15]

3.1.4 Ostatné hovorené databázy

Czech Speecon database je rozsiahlou databázou 550tich rečníkov. V súčasnej dobe však oficiálna webová stránka tohoto projektu nefunguje a k dátam sa nedá dostať ani z iných zdrojov.

CzechSpeechDat-E je databázou českej telefónnej komunikácie. Obsahuje nahrávky 1052 rečníkov (526 mužov a 526 žien). [16]

STAZKA obsahuje nahrávky z dopravných prostriedkov. Súbory sú dostupné na [12].

VYSTADIAL 2013/2016 je databázou obsahujúcou telefónnu komunikáciu, rovnako dostupná na [12].

Vzhľadom k vyššie uvedenému prehľadu českých hovorených korpusov, bol na základe požiadaviek na:

- kvalitu a spôsob nahrávania,
- vekovej rozmanitosti a pohlavnej vyváženosti rečníkov a
- voľnej dostupnosti nahrávok,

k vytvoreniu trénovacej databáze tejto práce vybraný korpus BUT – CZAS.

3.2 Spracovanie zvukových nahrávok

Po výbere vhodného rečového korpusu nasleduje jeho spracovanie. Pretože model strojového učenia má byť trénovaný na základe princípu učenia s učiteľom, je potrebné modelu na vstup dodať dáta, ktoré budú obsahovať informáciu o požadovanom výstupe. Požadovaný výstup modelu sú hodnoty prvých troch formantových kmitočtov. Cieľom tejto časti práce je túto databázu pripraviť.

Z nahrávok budú teda vypočítané hodnoty prvých troch formantov zaužívaným postupom odhadu formantových kmitočtov, ktorý bol popísaný v časti 1.3. Pre zaistenie čo najpresnejších hodnôt a eliminácie možnej chyby spôsobenej povahou tohoto postupu, bude každá hodnota počítaná tri krát, za využitia troch rôznych softvérov (Matlab, Praat a Wavesurfer).

K získaniu referenčných hodnôt formantov z nahrávok korpusu bol napísaný skript a niekoľko funkcií v prostredí Matlab. Účelom tohoto skriptu je automatizované vytvorenie celej databázy. Užívateľ potrebuje iba zvoliť parametre:

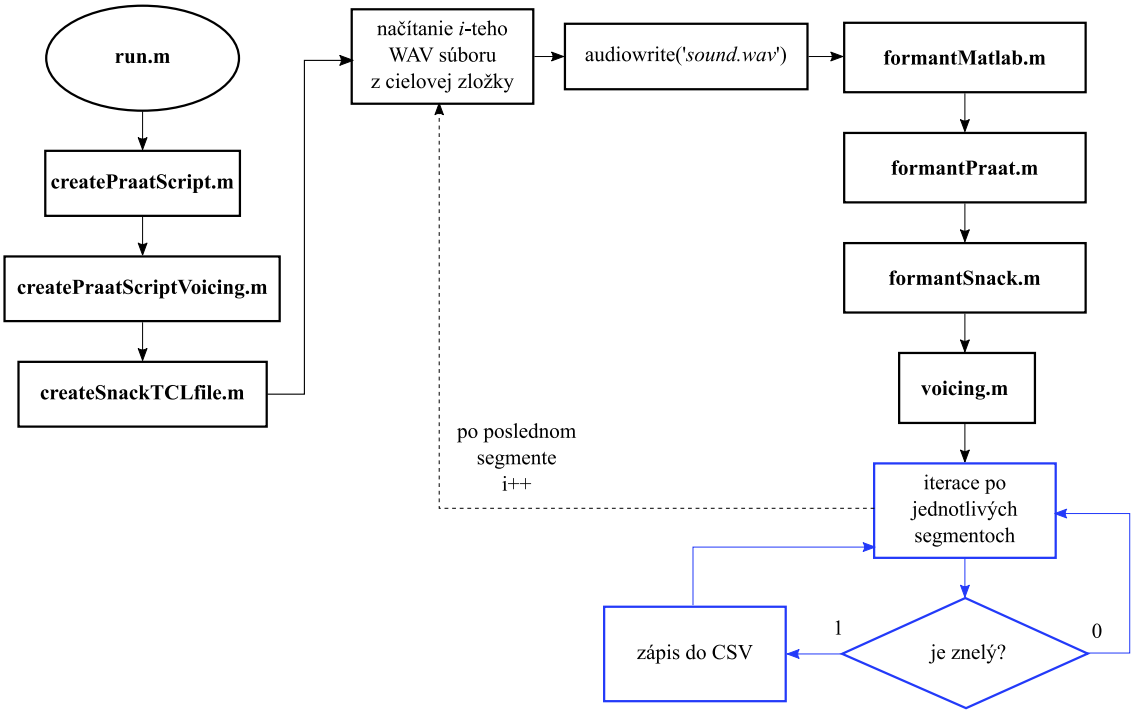
- w_L – dĺžka okna segmentu (Window Length),
- w_O – dĺžka prekyvu okien segmentov (Window Overlap)

a zadať cestu k zložke s zvukovými súbormi, ktoré sa majú spracovať. Pre účely tejto diplomovej práce, boli pre tieto parametre zvolené nasledujúce hodnoty:

- $w_\mathrm{L} = 25\,\mathrm{ms}$,
- $w_\mathrm{O} = 10\,\mathrm{ms}$.

Celý proces získania databázy referenčných hodnôt formantových kmitočtov je riadený zo skriptu `run.m`. Pre spustením tohoto skriptu má uživateľ možnosť zmeniť hodnoty w_L a w_O podľa vlastnej potreby. Rovnako môže zvoliť vlastný názov a umiestnenie CSV súboru do ktorého sa majú získané dáta uložiť. Po spustení skriptu je užívateľ vyzvaný k výberu zložky s nahrávkami, ktoré sa majú spracovať. Skript automaticky vyberie všetky súbory s príponou *.wav* ktoré sa v tejto zložke nachádzajú a postupne ich spracováva.

Na obrázku 3.1 je znázorný postup spracovania zvukových nahrávok riadený z Matlabu a postup volania jednotlivých `m`-file. Modrou farbou je označený úsek spracovania na úrovni jednotlivých segmentov.



Obr. 3.1: Bloková schéma spracovania zvukových nahrávok po spustení skriptu `run.m`.

K umožneniu priebehu všetkých častí tohoto skriptu a tým úspešnému vytvoreniu databáze formantových kmitočtov sú nevyhnutné tieto prerekvity:

1. Praat – aplikácia musí byť umiestnená v zložke, ktorá je v Matlabe namapovaná ako Current Folder (pri vytvorení bola používaná verzia Praat 6.0.43 z 8.9.2018) [17]
2. mPraat – všetky súbory a podzložky tohoto toolboxu musia byť pre Matlab dostupné - stačí ich pridať formou *Add to Path* (použitá verzia v1.1.3 z 20.10.2018) [18]
3. Tcl/Tk – musí byť naištalovaný v operačnom systéme (používané s verziou 8.5, od verzie 8.6 sa môžu vyskytovať problémy - upozorňuje na to aj webstránka Snacku) [19]
4. Snack – aktuálne použitá verzia 2.2.10 z 14.12.2005 Snack je iba toolkit, ktorý funguje pod Tcl, je preto potrebné Snack inštalovať až po úspešnej inštalácii Tcl (niektorých distribuciách Tcl je Snack už automaticky zaradený, iné vyžadujú samostatnú inštaláciu) [21]

3.2.1 Matlab

V tejto variante dochádza k výpočtu formantových kmitočtov priamo v programe Matlab. K tomuto účelu bol vytvorená funkcia v m-file `formantMatlab.m`. V prvom rade je signál predspracovaný – na celú nahrávku je aplikovaný preemfázový filter. Pretože formanty patria k segmentálnym príznakom, je pre samotným výpočtom kmitočtov potreba nahrávku segmentovať. Je teda volaná funkcia `segmentation.m`, ktorá vytvorí z nahrávky krátke úseky o žiadanej dĺžke. V rámci tejto funkcie sú jednotlivé segmenty násobené Hammingovým oknom.

Samotný výpočet hodnôt kmitočtov prebieha na základe modelovania spektrálnej obálky signálu pomocou lineárnej predikcie. Následne sú vybrané kladné korene prenosovej funkcie, z ktorých sú vypočítané hodnoty F_1, F_2 a F_3 podľa vzorca 1.3 a odpovedajúce šírky pásma B_1, B_2 a B_3 podľa vzorca 1.4. Skript výpočtu je napísaný podľa zdroju [20].

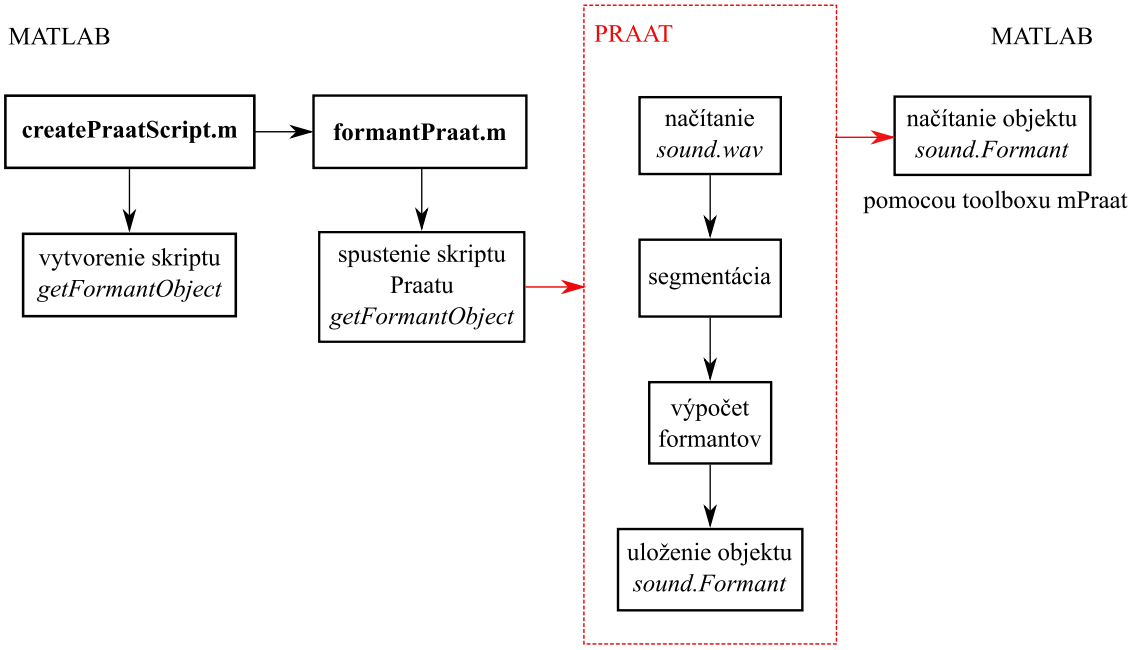
3.2.2 Praat

Praat je voľne dostupný počítačový softvér vyvíjaný od roku 1995 Paulom Boersmaom a Davidom Weeninkom na univerzite v Amsterdame. Ponúka rozsiahle možnosti analýzy a spracovania rečových signálov, zároveň s ich grafickým zobrazením, ale aj syntézu reči [17]. Samotný program je podporovaný pod rôznymi operačnými systémami a dá sa jednoducho používať z vlastného užívateľského rozhrania. Veľkou výhodou je možnosť vytvorenia vlastných skriptov pod koncovkou `.Praat`, ktoré sa dajú spúšťať samostatne z príkazového riadku. Toho je využité v tejto práci.

Aby bolo možné generovať databáze formantových kmitočtov z nahrávok automaticky, aj v prípade použitia rôznych hodnôt w_L a w_O, je súbor obsahujúci skript pre Praat vytvorený z Matlabu pomocou funkcie `createPraatFile.m`. Funkcia preberá hodnoty w_L a w_O, vytvára súbor *getFormantObject* v ktorom tieto hodnoty ukladá priamo do skriptu s príkazmi. Pomocou príkazu `system` je uložený skript v rámci programu Praat spustený na pozadí, rovnako ako z príkazového riadku. Predaním hodnôt je zabezpečené aby Praat nahrávku segmentoval s parametrami zadanými užívateľom v Matlabe.

Praat teda na pozadí extrahuje z nahrávky hodnoty formantových kmitočtov pre každý segment. Dáta uloží v súbore *segment.Formant*. Matlab potom tento súbor prečíta pomocou funkcií z toolboxu mPraat [18]. Celý proces je znázornený na obr. 3.2.

Pre vykonanie príkazov v skripte je nevyhnutné mať aplikáciu Praat uloženú v zložke, ktorá je namapovaná ako aktuálna zložka (Current Folder), v ktorej Matlab pracuje. V tejto zložke musí byť povolený zápis súborov, keďže výmena informácií medzi jednotlivými programami prebieha na základe ukladania súborov s dátami a



Obr. 3.2: Bloková schéma extrakcie formantových kmitočtov programom Praat.

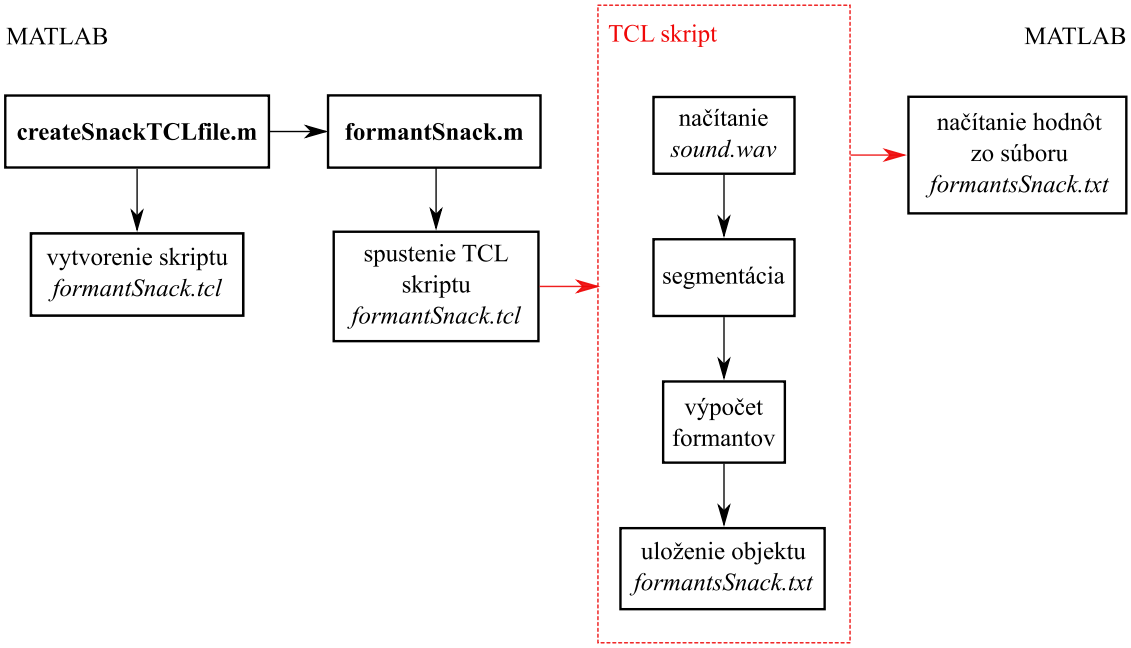
ich následného načítavania. Rovnako je potrebné aby bol do cesty, s ktorou Matlab pracuje pridaný adresár mPraat so všetkými podzložkami a súbormi. v tejto zložke umožnený zápis súborov, keďže výmena informácií medzi jednotlivými programmi prebbieha na základe ukladania a načítavania súborov v adresári.

3.2.3 Snack Toolkit (WaveSurfer)

WaveSurfer je aplikácia vyvinutá pre použitie v oblasti výskumu reči. Aplikácia je postavená s využitím toolkitu The Snack Sound Toolkit. Snack bol vyvinutý na Královskom technologickom inštitúte v Štokholme švédom Kåre Sjölander, najmä pre účely spracovania digitálnych nahrávok reči, ale ponúka aj jednoduchú vizualizáciu zvuku, ako napríklad spectrogram. Výhodou tohto toolkitu je možnosť kombinácie so skriptovacími jazykmi Python a Tcl/Tk. Táto kombinácia umožňuje vytváranie zvukových nástrojov a aplikácií s minimálnou námahou. Rovnaký skript sa dá potom použiť na rôznych operačných platformách. Samotná aplikácia WaveSurfer, rovnako ako celý Snack toolkit sú voľne dostupné z oficiálnej stránky projektu [21].

Pod jazykom Tcl/Tk je teda možné vytvoriť skript, s použitím príkazov z toolkitu Snacku. Tento skript sa dá potom jednoducho spustiť z príkazového riadku. Pomocou príkazu `system` je teda možné skript spustiť z prostredia Matlabu, rovnako ako v prípade skriptu programu Praat.

Na obr. 3.3 je schéma spracovania nahrávky a extrakcie formantov TLC skriptom s použitím knižnice Snack. Princíp je obdobný ako pri použití Praatu.



Obr. 3.3: Bloková schéma extrakcie formantových kmitočtov s použitím toolkitu Snack.

Na začiatku spustením `createSnackTCLfile.m` vytvorený samotný skript s príkazmi v jazyku Tcl, do ktorého uložíme aj hodnoty w_{L} a w_{O}, potrebné pre segmentáciu. V Matlabe funkcia `formantSnack` spustí tento skript na pozadí. Skript sa postará o načítanie aktuálnej nahrávky, jej segmentáciu na základe zadaných parametrov a následne vypočítané hodnoty prvých troch formantových kmitočtov a odpovedajúcich šíriek pásma uloží do súboru *formantsSnack.txt*. Matlab si informácie z tohoto textového súboru načíta a uloží v svojom prostredí pre možnosť ďalšieho spracovania.

Úspešne spustenie skriptu *formantSnack.tcl* predpokladá mať nainštalovaný a funkčný jazyk Tcl/Tk a rovnako nainštalovaný balíček Snack, ktorého príkazy skript k výpočtu formantov používa.

3.2.4 Znelosť

V kapitole 1 bolo vysvetlené, že formantová analýza má význam iba pre znelé časti reči. Preto je potrebné ešte vyriešiť otázku znelosti jednotlivých segmentov. Prakticky je znelosť riešená znova za použitia Praatu[1]. Praat totižto jedným príkazom

[1]predpokladá sa, že Praat ako sofistikovanejší softvér, priamo určený na analýzu rečových signálov dokáže presnejšie určiť znelosť úseku ako prípadný výpočet v Matlabe založený napríklad na porovnaní hodnoty krátkodobej energie a ZCR (počtu priechodov nulovou úrovňou)

Sound: To Pitch dokáže odhadnúť aktuálnu hodnotu F_0. K tomu je vytvorená funkcia `voicing` ktorá používa externý skript *getPitchObject* k odhadu F_0 pomocou Praatu, obdobne ako s formantovými kmitočtami v časti 3.2.2. Rozpoznávač znelosti potom v Matlabe funguje tak, že ak Praat pre segment nájde nenulovú hodnotu F_0, je segment prehlásený za znelý. V opačnom prípade, teda keď sa túto hodnotu nepodarí nájsť, je segment označený ako neznelý. Získaný vektor znelosti je následne filtrovaný jednorozmerným mediánovým filtrom za účelom vyhladenia.

3.2.5 Databáza

Pre každý formantový kmitočet boli odhadnuté tri hodnoty. Pre ďalší postup s databázou je ale potreba určiť jednu, čo možno najpresnejšiu hodnotu každého formantu. Toho je docielené na základe metriky jednorozmerného mediánového filtru, ktorý z týchto troch hodnôt vyberie vždy strednú hodnotu.

Do CSV súboru sú v tomto momente zapisované iba tie segmenty, ktoré boli označené ako znelé. Pre každý segment sú uložené základné parametre ako názov súboru z ktorého segment pochádza, poradové číslo prvého a posledného vzorku daného segmentu. Následne sú zapisované tri sady po troch formantových kmitočtoch – F_1, F_2 a F_3 – z každého z použitých softvérov. Doplnená je jedna sada stredných hodnôt z mediánového filtru. Hodnoty odpovedajúcich šíriek pásma B_1, B_2 a B_3 – sú v CSV súbory uložené za kmitočtami rovnakým spôsobom.

4 Vytvorenie modelu neurónovej siete

V tejto kapitole bude popísaný postup navrhnutie modelu neurónovej siete, jej trénovanie a evaluácia a konečné testovanie.

4.1 Extrakcia príznakov

Trénovanie neurónovej siete prebieha na základe dodania vektoru vstupných hodnôt a hodnôt očakávaných výstupov. Pri spracovaní rečových nahrávok by bolo možné na vstup použiť priamo pôvodné WAV dáta. Tento postup by však vyžadoval navrhnutie neurónovej siete s obrovským počtom vstupných neurónov a pokročilou architektúrou siete samotnej. Takýto návrh by bol zbytočne zložitý a mal vysoké výpočtové nároky. Preto je oveľa výhodnejšie vstupné dáta vybraným spôsobom parametrizovať a ako vstupné hodnoty použiť iba určitú číselnú reprezentáciu pôvodného signálu.

Formanty patria k segmentálnym príznakom a má význam analyzovať ich hodnoty výlučne pre každý segment zvlášť. Automaticky sa preto ponúka parametrizácia formou výpočtu určitého druhu koeficientov, ktoré samé o sebe dokážu vytvoriť vstupný vektor príznakov. K tomuto účelu sú vybrané lineárne predikčné koeficienty a melovské kepstrálne koeficienty. Obidve tieto reprezentácie signálu patria rovnako ako formanty medzi segmentálne príznaky.

4.1.1 Lineárne predikčné koeficienty

Princíp lineárnej predikčnej analýzy už bol načrtnutý v časti 1.3. Produktom lineárnej predikcie sú práve LPC koeficienty. Zo vzťahu 1.1 je zrejmé, že výsledný počet koeficientov a_k závisí na zvolenom ráde lineárnej predikcie p. Voľba hodnoty rádu modelu je dôležitou otázkou. V prípade, že je rád zvolený ako príliš malý, môže dôjsť k nedostatočnému predikovaniu charakteristiky. Opačný problém nastáva pri voľbe nevhodne vysokej hodnoty rádu predikcie. Doporučená hodnota závisí na použitom vzorkovacom kmitočte:

$$p = \frac{f_{\text{vz}}}{1000} + 2. \tag{4.1}$$

Pri ponechaní pôvodného vzorkovacieho kmitočtu nahrávok použitého korpusu $f_{\text{vz}} = 48\,\text{kHz}$ by bola doporučená hodnota $p = 50$ a koeficientov a_k by bolo rovnako 50 pre každý segment. Preto sú nahrávky pred výpočtom LPC koeficientov podvzorkované na novú hodnotu $f_{\text{vz}} = 16\,\text{kHz}$. Následne je generovaných 18 LPC koeficientov. V skripte `getLPC.m` sú implementované obidva tieto kroky.

4.1.2 Melovské kepstrálne koeficienty

MFCC – Mel Frequency Cepstral Coefficients sú prvé parametre, ktoré berú do úvahy nelinearitu ľudského sluchu spoločne s jeho maskovacími vlastnosťami. MFCC majú veľkú výhodu v tom, že ponúkajú dekorelovaný set parametrov, pretože je v postupe výpočtu využívaná diskrétna kosínusová transformácia, ktorá zabezpečuje že jednotlivé hodnoty nekorelujú [22]. Práve preto patria MFCC medzi úspešné a populárne príznaky pre využitie v trénovaní modelov strojového učenia.

Výpočet týchto koeficientov je založený na použití banky troujholníkových filtrov. Spektrum je bankou filtrov rozdelené lineárne v melovskej škále. Prepočet hodnôt frekvencí z Hz na mel prebieha vzťahom

$$f_{\text{mel}} = 2595 \cdot \log_{10}(1 + \frac{f_{\text{Hz}}}{700}). \tag{4.2}$$

S tým je spojený spätný prepočet z mel na Hz v tvare

$$f_{\text{Hz}} = 700 \cdot (10^{\frac{f_{\text{mel}}}{2595}} - 1). \tag{4.3}$$

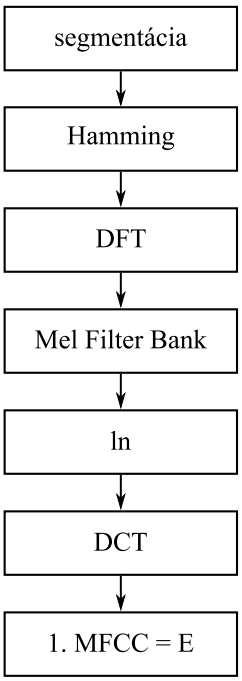
Spätným prepočtom je docielené nelineárne rozloženie filtrov v mierke Hz. Táto skutočnosť však vyhovuje účelom použitia v spojení s formantmi. Banka filtrov má hustejšie rozdelenie na nižších kmitočtoch, kde sa okrem väčšiny užitočných informácií pre spracovanie reči nachádzajú aj formanty. Vyššie kmitočty sú v redšom zastúpení, čo sa hodí, pretože pri analýze prvých troch formantových hodnôt nie sú informácie obsiahnuté v spektre na vysokých kmitočtoch dôležité.

MFCC patrí k segmentálnym parametrom signálu, preto sa počíta pre každý segment zvlášť. Výpočet MFCC teda umožňuje pomerne jednoducho získať set parametrov, kompaktne reprezentujúcich spektrum každého segmentu, ktorý sa dá využiť ako set vstupných dát pre trénovanie modelu strojového učenia.

Postup výpočtu koeficientov sa skladá z niekoľkých po sebe nasledujúcich krokov, viď obr. 4.1 [23]. Rovnako ako u LPC je signál najprv podvzorkovaný na 16 kHz. Táto hodnota f_{vz} bude odpovedať ideálnemu počtu 20tich počítaných koeficientov. Podvzorkovanie aj samotný výpočet koeficientov MFCC je riešený vo fukncii `mfcc.m`, ktorá volá skript `melBank` k vytvoreniu banky troujholníkových filtrov.

4.2 Návrh a trénovanie modelu

Návrh modelu neurónovej siete a jej trénovanie prebiehalo v prostredí Matlab za pomocou nástrojou `nnstart` a `nntool`. Trénovaný bol model neurónovej siete. Keďže cieľom je z každého segmentu predikovať 3 rôzne hodnoty formantových kmitočtov, je možnosť hneď na začiatku vybrať z dvoch od základu rozdielnych návrhov.



Obr. 4.1: Bloková schéma výpočtu MFCC koeficientov.

Prvá možnosť je pre každý z troch formantov natrénovať samostatnú neurónovú sieť. Každý model by v tomto prípade mal zhodný počet neurónov na vstupnej vrstve, závislý na počte prvkov vstupného vektoru príznakov. Následne voľne voliteľný počet skrytých vrstiev s ľubovoľným počtom neurónov v každej vrstve. Na výstupe každého z týchto modelov by musel byť práve jeden neurón. Každý model by predstavoval bežný regressor s jedným výstupom. Ako výhoda by sa naskytla možnosť navrhutia rozdielnych rozmerov siete pre každý z formantov, tak aby mu bola ušitá na mieru. Nevýhodou však zostáva že jednotlivé regresory nebudú schopné zdielať informácie medzi sebou a navzájom sa ovplyvňovať.

Druhá možnosť je navrhnúť jednu neurónovú sieť s 3 neurónmi vo výstupnej vrstve. Tento tvar nie je neobvyklý v prípade klasifikácie, kde 3 výstupné neuróny predstavujú 3 klasifikačné skupiny. Regresný problém s viacerými výstupmi súčastne sa nazýva *multi-target regression* alebo *multi-output regression* a nie je zas tak bežný. Principiálne by ho však neurónové siete mali zvládnuť.

V modeloch vytvorených pomocou `nnstart` nie je možné zvýšiť počet vrstiev skrytých neurónov na viac ako jednu. Rovnako je pevne preddefinovaná aktivačná funkcia skrytej vrstvy ako logistická funkcia a výstupnej vrstvy ako lineárna funkcia. Použitá sieť je typu feed-forward, učenie prebieha na princípe backpropagation, teda spätného šírenia chyby pomocou Levenberg-Marquardtového algoritmu.

Zostáva možnosť experimentovať s počtom neurónov v skrytej vrstve. Zároveň dispozícia dvoch sád príznakových parametrov z predchádzajúcej kapitoly ponúka

alternáciu až 3 rôznych skupín vstupných vektorov - LPC koeficienty, MFCC koeficienty alebo obidva dohromady.

Prvé tri formantové kmitočty naberajú rôzne hodnoty, takže sa pohybujú v rôznych rozsahoch. Navyše tieto rozsahy sú pre každý formant rozlične veľké. To by mohlo mať nepriaznivý vplyv na učenie modelu. Dokumentácia Matlabu v prípade regresie s viacerými výstupmi, ktorých hodnoty sú v rozdielnych rozsahoch, doporučuje normalizáciu hodnôt výstupných dat. Rozdielne rozsahy hodnôt MSE na jednotlivých výstupoch by mohli spôsobiť snahu modelu znižovať absolútne najväčšiu chybu, nie tú relatívnu. Preto sú hodnoty formantov pred samotným trénovaním modelu normalizované pre každý z formantov zvlášť, tak aby všetky formanty obsahovali hodnoty od 0 do 1 a boli si číselne vzájomne rovnocenné. Je potrebné uložiť maximálne a minimálne hodnoty formantom, ktorými je dataset normalizovaný, aby bolo možné predikované hodnoty spätne prepočítať do pôvodného rozsahu.

Databáza vytvorená v rámci tejto práce obsahuje takmer 600 000 prvkov. Je zbytočné a časovo až kontraprodutívne vo fázi návrhu modelu siete používať tak veľké množstvo prvkov. Množina je preto v tejto fázi znížená na 20tinu, vybraním každého 20. prvku, s predpokladom, že prípadné rozšírenie trénovacej skupiny dokáže dosiahnuté výsledky ešte zlepšiť.

Zároveň námatkovo oddelím od trénovacej množiny 500 prvkov, ktoré sa na trénovaní nebudú podielať vôbec. Týmto datasetom bude najúspešnejší model testovaný v závere.

Za týchlo predpokladov bolo trénovaných viacero modelov s rôznym počtom neurónou v skrytej vrstve. Najčastejšie bolo použitých 9, 25 alebo 40 neurónov. Na mieste vstupných príznakov boli striedané už spomínané tri skupiny vektorov, tak aby bola nájdená čo najlepšia možná kombinácia s najmenšou chybou MSE. Bol odpozorovaný trend toho, že modely s MFCC koeficientami alebo kombináciou MFCC a LPC koeficientov boli úspešnejšie ako modely za použitia LPC samotných (často ale iba o niekoľko tisícin). MFCC a kombinácia obidvoch boli často rovnako úspešné, ale dvojnásobné množstvo vstupných dát u kombinácie týchto koeficientov vyžadovalo väčšie nároky na výpočet a samotný výpočet spomalilo. Rozdiel v úspešnosti pritom nebol tak veľký aby sa toto zaťaženie vyplatilo. Pri použití pôvodného veľkého datasetu by spomalenie ešte narástlo niekoľkonásobne.

Na základe toho bol ako najvhodnejší z navrhovaných modelov vybraný model s MFCC koeficientami na vstupoch a počtom 25 neurónov v skrytej vrstve. Tento model dosiahol pri trénovaní MSE = 0,0024 na trénovacej množine a 0,0026 na množine validačnej.

Následné pokusy o zníženie tejto chyby pridaním daľšej skytej vrstvy neutrónov a upravením ich počtu neboli úspešné. Výsledná MSE buď bola veľmi podobná ale trénovanie trvalo dlhšie, alebo sa výsledná chyba stúpla.

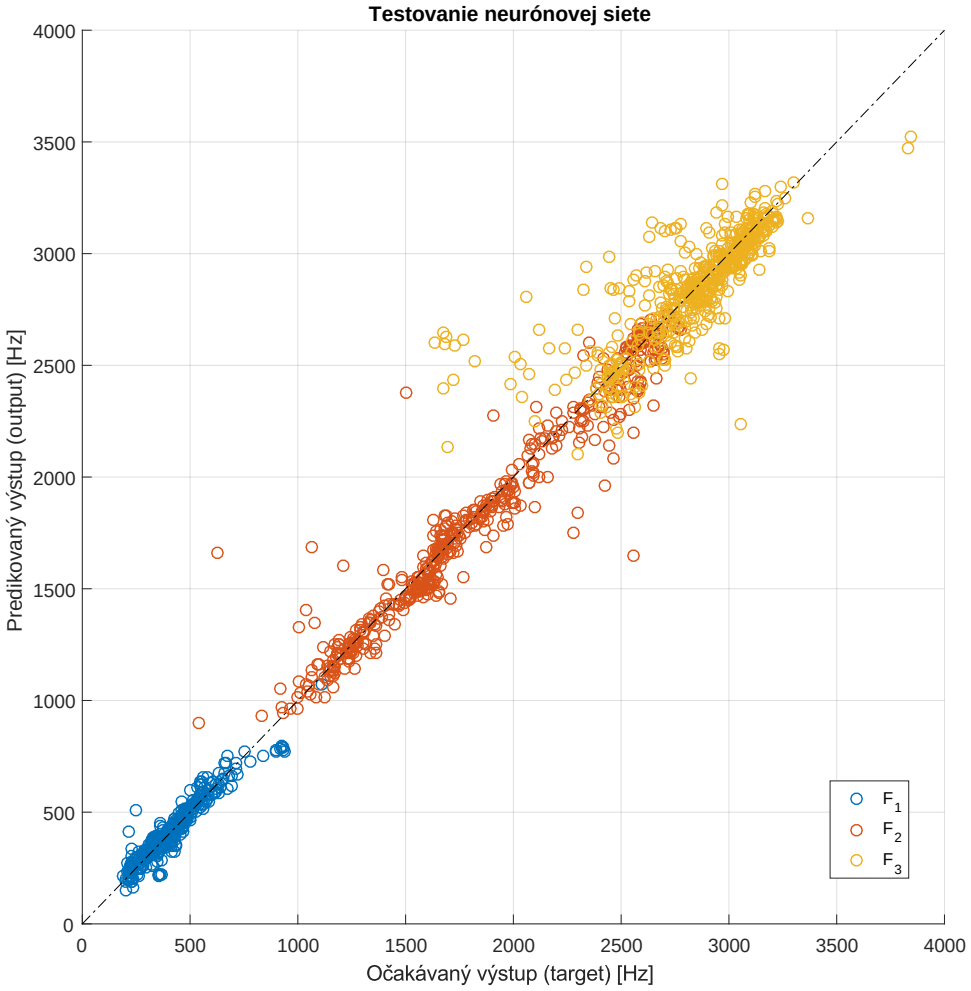
Tento návrh bol následne opätovne, tento krát boli k trénovaniu dodané všetky prvky datasetu (okrem množiny vyhradenej k záverečnému testovaniu). Rozšírenie počtu trénovacích prvkov dokázalo chybu na validačnej množine mierne znížiť. Pri použití veľkého datasetu bolo pri trénovaní dosiahnuté chyby MSE = 0,0024 na trénovacej a rovnako na validačnej množine. Natrénovaný model tejto siete bol uložený a ďalej testovaný.

4.3 Testovanie modelu

Pre účely záverečného testovania vytvoreného modelu bolo pred fázou trénovania vytýčených 500 vzorkov z databázy, ktoré sa na trénovaní nepodieľali. Hodnoty na výstupe boli týmto modelom predikované a následne porovnané s očakávanými hodnotami. Chyby sú zobrazené v tabuľke 4.1, grafické porovnanie s rozdelením jednotlivých formantov na obrázku 4.2.

Tab. 4.1: Testovanie modelu

	Chyby po počítané na normalizované hodnotách	Hodnoty po spätnej normalizácií
MSE	0,0014	17 604
RMSE	0,0369	132,68
MAE	0,0203	69,83
MAPE	8,24 %	5,51 %



Obr. 4.2: Graf porovnania predikovaných a očakávaných výstupov jednotlivých formantov.

5 Záver

Táto diplomová práca rozoberá problematiku odhadu formantových kmitočtov. V prvej časti tejto práce je rozoberané teoretické pojednanie o problematike formantov a načrtnutie možností využitia niektorých algoritmov strojového učenia.

Jedným z vytýčených cieľov zadania bolo vytvorenie referenčnej databáze vzoriek znelej reči pomocou softvérov Praat a WaveSurfer. Tento cieľ bol v práci plne splnený. V prostredí Matlab bol za týmto účelom vytvorený systém funkcií, ktoré sú schopné zavolať tieto dva softvéry s pomocou externých skriptov počítať hodnoty prvých troch formantových kmitočtov automatizovane pre všetky zvukové súbory v označenej zložke. Užívateľ si pritom môže ľubovoľne zmeniť prednastavené hodnoty veľkosti okna segmentu a veľkosti prekrytia okien.

Referenčná databáza vzoriek znelej reči bola vytvorená s dĺžkou okna 25 ms a veľkosťou prekrytia 10 ms. K týmto hodnotám očakávaných výstupov boli pridané LPC a MFCC koeficienty, ktoré plnia úlohu vstupných dát. V prostredí Matlab bol navrhnutý model neurónovej siete. Tento model bol trénovaný prvkami databáze. V závere bol model testovaný na testovacej množine prvkov. Výsledná percentuálna chyba predikcie sa rovná 5,51 %. Záverečný graf znázorňuje presnosť predikovania formantových kmitočtov. Z tohoto grafu je možné jednoducho vypozorovať, že model odhaduje lepšie kmitočty formantov na nižších hodnotách, najmä F_1.

Literatúra

[1] KIM, CH.; K. SEO a W. SUNG. A Robust Formant Extraction Algorithm Combining Spectral Peak Picking and Root Polishing. *EURASIP Journal on Advances in Signal Processing.* [online]. 2006, 2006(67960), s. 1-16. ISSN 1687-6180.

[2] KRČMOVÁ, M. *Fonetika a fonologie* [online]. 3 vyd. Brno: Masarykova univerzita, 2009 [cit. 2018-11-16]. Elportál. Dostupné z: <`http://is.muni.cz/elportal/?id=852835`>. ISSN 1802-128X.

[3] PSUTKA, J.; et al. *Mluvíme s počítačem česky.* Praha: Academia, 2006. Česká matice technická (Academia). ISBN 80-200-1309-1.

[4] SMÉKAL, Z. *Číslicové zpracování řeči* [elektronická skripta]. Brno: VUT v Brně, 2010 [cit. 2018-11-22].

[5] PALKOVÁ, Z. *Fonetika a fonologie češtiny.* Praha: Karolinum, 1994. ISBN 80-7066-843-1.

[6] SKARNITZL, R.; VOLÍN, J. Referenční hodnoty vokalických formantů pro mladé dospělé mluvčí standardní češtiny. *Akustické listy.* FF UK: ČsAS, 2012, 18(1), 7-11. Dostupné z: <`https://fonetika.ff.cuni.cz/wp-content/uploads/sites/104/2016/05/Ska_Vol2012.pdf`>

[7] HONZÍK, P. *Strojové učení.* [elektronická skripta]. Brno: VUT v Brně, 2006.

[8] MACHOVÁ, K. *Strojové učenie: princípy a algoritmy.* [elektronická skripta]. Košice: TU Košice, 2002.

[9] NÁVRAT, P.; BIELIKOVÁ, M.; BEŇUŠKOVÁ, L.; et al. Umelé neurónové siete. *Umelá inteligencia.* Bratislava: Vydavatel'stvo STU, prvé vydanie, 2002, ISBN 80-227-1645-6, 157–196 s.

[10] CHEN, T.; GUESTRIN C. XGBoost: A Scalable Tree Boosting System. *roceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* San Francisco, California, USA, 2016, KDD '16(2939785), 785–794.

[11] *Český národní korpus* [online]. Praha [cit. 2018-12-12]. Dostupné z: <`https://www.korpus.cz/`>

[12] *LINDAT/CLARIN* [online]. Praha: UFAL MFF UK, 2018 [cit. 2018-12-12]. Dostupné z: <`https://lindat.mff.cuni.cz/cs/`>.

[13] HÁJEK, V.; HARAR, P.; SCHIMMEL, J.; et al. BUT-CZAS: Korpus kvalitních nahrávek české řeči pořízených v bezodrazové komoře. *Elektrorevue*. 2018, 20(2), 48–50. ISSN 1213-1539. Dostupné z: <http://splab.cz/download/databaze/but-czas>

[14] Vytváření řečových korpusů. *Katedra kybernetiky ZČU* [online]. Plzeň: ZČU, 2019 [cit. 2019-05-13]. Dostupné z: <http://www.kky.zcu.cz/cs/research-fields/speech-corpus-creation>

[15] Audiovizuální korpus UWB-05-HSCAVC. *Katedra kybernetiky ZČU* [online]. Plzeň: ZČU, 2019 [cit. 2019-05-13]. Dostupné z: <http://www.kky.zcu.cz/cs/research-fields/audio-visual-corpus-UWB-05-HSCAVC>

[16] SpeechDat-E www home page. *Fakulta elektrotechniky a informatiky* [online]. Brno: VUT, 2001 [cit. 2019-05-13]. Dostupné z: <http://www.fee.vutbr.cz/SPEECHDAT-E/>

[17] BOERSMA, P.; WEENINK, D. *Praat: doing phonetics by computer* [Počítačový program]. 2018, ver. 6.0.43, aktualizované 8.9.2018 [cit. 2018-12-12]. Dostupné z: <http://www.praat.org/>.

[18] BOŘIL, T.; SKARNITZL, R. Tools rPraat and mPraat. *Text, Speech, and Dialogue: 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings*, 2016; s. 367 – 374. Dostupné z: <https://fu.ff.cuni.cz/praat/>

[19] Tcl/Tk 8.5. *Tcl Developer Site* [online]. [cit. 2019-05-16]. Dostupné z: <https://www.tcl.tk/software/tcltk/8.5.tml>.

[20] Formant Estimation with LPC Coefficients – MATLAB & Simulink. *MathWorks – Makers of MATLAB and Simulink* [online]. 2018 [cit. 2018-12-12]. Dostupné z: <https://www.mathworks.com/help/signal/ug/formant-estimation-with-lpc-coefficients.html>

[21] SJÖLANDER, K. Snack Home Page. *Division of Speech, Music and Heering* [online]. Stockholm: KTH, 2006 [cit. 2019-05-05]. Dostupné z: <http://www.speech.kth.se/snack/>

[22] HOSSAN, A.; SHEERAZ, M. a M. A. GREGORY. A novel approach for MFCC feature extraction. *2010 4th International Conference on Signal Processing and Communication Systems.* IEEE, 2010, s. 1 – 5.

[23] MUDA, L.; BEGAM, M. a I. ELAMVAZUTHI. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *Journal Of Computing.* 2010, 2(3), s. 138 – 143. ISSN 2151-9617.

Zoznam symbolov, veličín a skratiek

a_k	LPC koeficienty
B_i	šírka pásma i-teho formantu
F_0	kmitočet základného tónu
F_i	i-ty formant
f_{vz}	vzorkovací kmitočet
LPC	Linear Predictive Coding – Lineárne predikčné kódovanie
MFCC	Mel Frequency Cepstral Coefficients – Melovské kepstrálne koeficienty
MSE	Mean Square Error
p	rád lineárnej predikcie
T_0	perióda základného tónu
w_{L}	Window Length – dĺžka okna (segmentu)
w_{O}	Window Overlap – dĺžka prekryvu okien (segmentov)

Zoznam príloh

A Obsah priloženého CD

Na priloženom CD sa nachádza elektronická verzia tejto diplomovej práce. Ďalej sa tu nachádzajú všetky všetky skripty vytvorené v rámci tejto práce.